DOCUMENT RESUME

ED 270 498 TM 860 388

AUTHOR Mandeville, Garrett K.
TITLE The Evaluation of a "Count Down" Procedure for Identifying Effective Schools.
PUB DATE 17 Apr 86
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Gains; Achievement Rating; Achievement Tests; Awards; Educational Improvement; Elementary Secondary Education; *Evaluation Criteria; Evaluation Methods; *High Achievement; Improvement Programs; Institutional Evaluation; Mathematics Achievement; Reading Achievement; *Research Problems; Sampling; *School Effectiveness; State Programs; Validity
IDENTIFIERS Comprehensive Tests of Basic Skills; *Effective Schools Research; South Carolina; South Carolina Basic Skills Assessment Program; South Carolina Educator Improvement Act; *South Carolina School Incentive Reward Program

ABSTRACT

South Carolina's School Incentive Reward Program Program (SIRP) based on year-to-year gains in reading and mathematics achievement, which seemed to have face validity, was evaluated, particularly its use of the relative gains of students in only the highest grade in the school as the basis for the awards. When this requirement was relaxed so that the results of the students in the next lower grade in the school were used, award decisions based on the two analyses were inconsistent, essentially random. The data were obtained from South Carolina's Basic Skills Assessment Program's (BSAP) criterion referenced tests administered to all students in grades 1-3, 6, 8, and 11, and the Comprehensive Tests of Basic Skills (CTBS) administered in grades 4, 7, and 10. The percentage passing the BSAP tests and the school mean normal curve equivalents on the CTBS subtests defined school-grade performance. Schools with high achievement gains or generally high performance were considered eligible to receive SIRP funds. Of 984 schools studied, 315 were identified as high performance schools based on the highest grade level, and 310 based on the next lower grade. One hundred and three schools were identified on both analyses. It was suggested that policymakers be aware of these problems in effective schools research. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270498

The Evaluation of a "Count Down" Procedure for Identifying Effective Schools

Garrett K. Mandeville

University of South Carolina

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. K. Mandeville

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 860 388

Paper presented at the annual convention of the American Educational Research Association, San Francisco, April 17, 1986.

Abstract

A proposed school incentive reward program based on year-to-year gains in reading and mathematics achievement, which seemed to have "face validity", was evaluated. The salient feature of the proposal which was scrutinized was the utilization of the relative gains of students in only the highest grade in the school as the basis for the awards. When this requirement was relaxed so that the results of the students in the next lower grade in the school were used, award decisions based on the two analyses were inconsistent, essentially random. It is suggested that the results are due to either methodological problems (especially the nature of the data) or to inherent inconsistency of improvement across grades within a school. It is important that policy makers are aware that many unsolved problems still exist in the identification of truly "effective schools."

In 1984 the South Carolina legislature passed comprehensive legislation called the Education Improvement Act (EIA). Funded by an increase in the state sales tax from 4% to 5%, main components of the act provided funds for increasing teacher salaries, remediating underachieving students, and financing school building programs. Other important provisions of the EIA provided funds for school, principal, and teacher incentive awards. This paper deals only with procedures related to the identification of schools to receive the approximately $7 million earmarked for the School Incentive Reward Program (SIRP).

The EIA provided guidelines for the eligibility of schools to receive SIRP funds which were, in a sense, quite specific. The legislation required that schools must exhibit "exceptional performance for criteria such as: (a) achievement gains over prior year..." and that "No school or district is eligible for incentive grant funds unless the achievement gain criterion is met." (EIA, 1984) Increased achievement then, based on a year-to-year (i.e., Spring, 1984 to Spring 1985) comparison, was a necessary (but not sufficient) condition for eligibility. Although not the only condition, increased, increased achievement clearly dominated the other eligibility criteria which had to do with teacher and student attendance. (In fact, the latter criteria were eventually used only to determine the size of the reward not whether one was given.) The legislation was quite general, however, in that no direction was provided on how year-to-year school achievement gains might be operationalized.

A Departmental Proposal

The writer became involved by being asked to evaluate a procedure for identifying eligible schools which was proposed by the division within

the agency which had responsibility for this task. The proposal called for the comparison of achievement in reading and mathematics of students at a particular grade level based on Spring 1983 and Spring 1984 testing periods. (Since the procedure was being evaluated, 1983-1984 comparisons were to be considered rather than the 1984-1985 comparisons upon which awards were actually to be made.) A brief summary of the proposal follows.

The proposal had three defining characteristics. The first was that schools would be identified as eligible (hereafter to be termed as "identified") based on what might be termed "exceptional" gains or, if the school was already achieving at a relatively high level, based on more modest gains. The former group of schools will be termed as "high gain" schools and the latter as "high performance" schools. Second, high gain and high performance were to be defined in relative rather than absolute terms. Year-to-year gains were to be cast into a distribution and some percentage, e.g., the top 40% were to be considered as "high gain." (The percentages could be adjusted, depending upon the results.) The third characteristics of the proposal dealt with the question of which grade or grades from each school should be utilized in the analysis. (Statewide test data were available at all but two grade levels in the 1-11 grade span.) The proposal called for the data from the highest grade only (for which, of course, statewide data were available) to be used to "measure" changes in the achievement level of each school.

Although the reader might quibble with certain aspects of the proposal, (e.g., its focus on the basic skills which is probably less appropriate for secondary than elementary schools, the use of a single

grade from a school which is probably more appropriate for a school containing only grades 6 through 8 rather than one with a K-8 organizational structure) it seems to have a certain amount of "face validity." The primary charge to the writer was to develop a computerized system to implement the proposal and during the process to "adjust" the definitions of "high gain" and "high performance" so as to provide information on the number of schools identified for each definition. In addition to this primary responsibility, the effect of relaxing the requirement that the highest grade in the school be utilized was to be investigated. In this paper, the results of this secondary study are reported.

Theoretical Perspectives

The identification of schools to be eligible for SIRP rewards is, to some extent, related to the "effective schools" research efforts. If we can simplify this research area to 1) the identification of schools where "something is happening", and 2) the study of those schools where "something is happening" in an attempt to determine "why it is happening" then the legislature clearly was more interested in the former than the latter. It is an understatement to say that there are many unresolved questions surrounding the identification of effective schools. Various approaches based on absolute gains, trends, and regression methods tend to produce inconsistent results (e.g., Frechtling, 1982) and different regression based methodologies do not even agree (Dyer, Linn and Patton, 1969). Furthermore, many of these studies have not dealt with the question of what might be called "the grade level problem." In fact, a common criticism of the early school effectiveness studies was the utilization of test data from only one grade level.

Concerning this issue, Ralph and Fennessey (1983) recommended using data from more than one grade and Rowan, Bossart and Dwyer (1983) suggested that all grades be considered. These researchers obviously do not believe that school main effects are so dominant and pervasive as to be reliably identified on the basis of probing a school at a single grade level. Another major issue in the assessment of school effects has to do with the use of data from successive classes at a given grade level rather than longitudinal data. Use of the former rather than the latter (as was done in this study) is usually viewed as the practical but inferior alternative. The situation is so confused that one of the foremost researchers in the field (Rowan, 1985) has recently expressed the opinion that "The best method of measuring school effectiveness is unknown" (p. 99).

Methods

The data were obtained from two statewide testing programs which have been in effect in South Carolina for a number of years. The Basic Skills Assessment Program (BSAP) employs criterion referenced tests of all students at grades 1-3, 6, 8 and 11; all students in grades 4, 7 and 10 are tested with the Comprehensive Test of Basic Skills (CTBS). Thus, in the 1-11 grade span, statewide testing occurs at all but the fifth and ninth grades. Percent passing the reading and mathematics BSAP tests and school mean Normal Curve Equivalents (based on the Total Reading and Total Mathematics subtests on the CTBS) were the operational definitions of school-grade performance. At each grade level 1983 to 1984 gains were case into a distribution and the percentile ranks of these gains were determined. In this way, the problem of the lack of consistency of the two metrics was overcome. The school-grade records were than merged by school with the record for the highest available

grade in the school being the only one retained. High gain schools were defined as those with percentile ranks of 60 or greater for both reading and mathematics gains.

In order to identify high performance schools, the percentile ranks of the 1983 measures were also included in the original grade-specific data sets. At each grade level, those schools achieving at or above the 60th percentile in both reading and mathematics were subset and the median gains (in the appropriate "raw" metric) in reading and mathematics were obtained for these school-grade subsets. These median gains (or zero in the event that they were negative) were then included in the merged (by school) data subset of high performing schools. Finally, high performance schools were defined as those in the high performing group which exhibited gains in both subjects that exceeded the median gain (or possibly zero).

In keeping with the original proposal, the highest grade (for which data were available) was used to identify schools eligible for the award. In a second analysis, the next lower grade for which data were available (if one existed) was used as the grade upon which the selection decision was based. The consistency of these two sets of results was then considered as a function of grade level and instrument consistency (i.e., whether the decisions were based on different tests).

Data Sources

As noted above, the BSAP tests are criterion referenced and they were used at grades 1-3, 6, 8 and 11. They contain 36 reading and 30 mathematics items (except at grade 11 where the reading and mathematics tests contained 60 to 50 items respectively), all multiple choice. The appropriate levels of Form U of the CTBS (90 reading and 85 mathematics items) were administered at grades 4, 7 and 10.

Results

After editing and merging the data for the two years, 1032 schools were represented on the data base. For 48 schools, however, the grade organizational structures were such that statewide data existed for only one grade so they were excluded from the analyses. For the remaining 984 schools, 315 were identified according to the first analysis (using the penultimate grade) and 310 were identified when the next lower grade was used. The number of schools identified on both analyses was 103. The coefficient of decision consistency, kappa, was a modest .018 which is about one-half of its null-model standard error using a formula provided by Fleiss, Cohen and Everitt (1969). The results by "highest grade" are presented in Table 1.

Table 1 about here

We note that, with the exception of grades 6 and 11, the magnitude of kappa is less than its null standard error, usually much less. (Since the standard error formula is based on assymptotic theory, the results are somewhat suspect for grades 2 and 7.) It is interesting to note that, although the kappa values do not approach normal standards for statistical significance for either grade 6 or grade 11, the estimates are of opposite sign. At the risk of over-interpreting nonsignificant results, the negative 6th grade estimate suggest that at least some schools focused improvement efforts on either the 4th or 6th grade in 1984 with the extra resources needed for the selected grade leading to lesser performance (not necessarily a decrease but at least not as large as increase) for the nonselected grade. Thus the

"inconsistency" would be slightly greater than expected by chance. A large majority (89%) of the schools included in the 6th grade analysis were organized along traditional elementary school lines, containing grades 1-6. Possibly principals in such schools are more sensitive to problems which may exist in certain grades and reallocate resources to deal with these problems. If this interpretation is correct, it runs counter to the belief that effects in elementary schools should be more consistent than those in secondary schools because of the impact of the principal as the instructional leader in the smaller organizational setting. The (again nonsignificant but positive) kappa in grade 11 was primarily based on schools with fewer grade levels, either 9-12 or 10-12. It is possible that improvement efforts in such an environment might be more general in nature and, therefore, be reflected more consistently across the grades involved.

Although selection decisions were generally quite inconsistent, we might consider how they related to the compatibility of the two types of tests involved. The amount of data available for the BSAP-BSAP comparisons is limited to the 2nd and 3rd grades where the kappas were small relative to their standard errors. The two grade levels discussed above (grades 6 and 11), for which the kappas at least exceeded their standard errors in magnitude, were based on BSAP-CTBS comparisons. (In this notation, the first test listed refers to the test used in the higher grade.) The kappas for the other BSAP-CTBS comparison (grade 8) and for the two CTBS-BSAP comparisons were very small, however. In summary, it does not seem that consistency was related to whether the test at the substitute grade level was of the same or a different type. Inspection of the results further disaggregated to reflect school organizational structure provided no new insights.

Educational Significance

If one takes the viewpoint that schools vary on factors which affect achievement such as quality of instruction, that this variation will rather consistently manifest itself across grades, and that these factors directly impact on gains in student achievement, the original proposal provides a reasonable approach to the problem of identifying exceptional schools. The negative findings indicate that, if the first assumption is true (which is almost axiomatic), then either the second or the third must be false. The fact that the BSAP tests were developed based on statewide educational objectives, that they are psychometrically sound (albeit, quite short), and that the inconsistency of results was so uniform across the factors considered, suggest that instrumentation to support the third assumption may have been adequate. If this is the case, we are forced to conclude that either the methodology used was insufficient to identify school effects or that school effects may, in large part, actually be "grade level within school" effects. A primary deficiency with the methodology (but surely not the only one) was the reliance on comparisons of the performance of different groups of students. Given this constraint, however, it is clear that in the evaluation of schools for incentive rewards, the total school should be considered. Even then, large aggregate indices will likely mask grade levels where performance is less than exceptional. Finally, if studies employing longitudinal data tend to verify these results, this would suggest that much of the early effective schools research has been misguided. It is important that policy makers and those who advise them be made aware of the problems associated with the identification of exceptional schools.

Table 1

Statistics on Decision Consistency

High Grade	Next Lower	Percentages ++	(+-/-+)	--	Total N	Kappa	St. Error
2	1	23.8	42.9	33.3	21	.152	.243
3	2	6.1	33.8	55.1	49	-.020	.144
4	3	9.2	42.2	48.6	294	.002	.059
6	4	6.3	44.7	49.0	190	-.093	.073
7	6	10.4	44.8	44.8	29	-.005	.201
8	7	10.7	42.9	46.4	205	.017	.070
11	10	15.8	41.3	42.9	196	.110	.072
All		10.5	42.6	46.9	984	.018	.032

Note: "++" indicates that a school was identified on both the first and second analyses, "--" that a school was not identified on either, and "+-" and "-+" represent inconsistently identified schools.

References

Dyer, H. S., Linn, R. L. and Patton, M. J. (1969) A comparison of four methods of obtaining discrepancy scores based on observed and predicted school system means on achievement tests. American Educational Research Journal, 6, 591-605.

Education Improvement Act, S. C. Code of Laws 1976, (as amended) 59-18-10 (1984).

Fleiss, J. L., Cohen, J. and Everitt, B. S. (1969) Large sample standard errors of kappa. Psychological Bulletin, 72, 323-327.

Frechtling, J. A. (1982) "Alternative methods for determining effectiveness: Convergence and divergence." Paper presented at the annual meeting of the American Educational Research Association, New York.

Ralph, J. H. and Fennessey, J. (1983). Science or reform: Some questions about the effective schools model. Phi Delta Kappan, 64, 689-694.

Rowan, B. (1985. The assessment of school effectiveness. In Kyle, R. M. J. (Ed.) Reaching for excellence: An effective schools sourcebook. N.I.E.: Washington, D. C.

Rowan, B., Bossert, S. T. and Dwyer, D. C. (1983). Research on effective schools: A cautionary note. Educational Researcher, 12, 24-31.